Correction to Gerber and Green (2000), Replication of Disputed Findings, and Reply to Imai (2005)

ALAN S. GERBER and DONALD P. GREEN *Yale University*

This essay corrects the results reported in Gerber and Green 2000 and replies to Imai (2005). When data-processing errors are repaired, the original substantive findings from the New Haven experiment remain unchanged. As previously reported, brief phone calls do not increase voter turnout. The "correction" that Imai (2005) offers, which purports to show that these phone calls produce large, significant, and robust increases in voter turnout, is shown to contain statistical, computational, and reporting errors that invalidate its conclusions about the effectiveness and cost-effectiveness of phone calls and mail. A replication of the New Haven experiment reproduces both the findings reported in Gerber and Green 2000 and the biases of Imai's statistical analysis.

We are grateful to Kosuke Imai, who, in the course of replicating and extending our experimental analysis (Gerber and Green 2000), brought to our attention data-processing problems associated with our 1998 experiments on voter mobilization. However, the "correction" of our findings reported in Imai 2005 contains numerous statistical, computational, and reporting errors. The purpose of this essay is twofold: to offer a correction of our original findings and to demonstrate that the statistical analysis conducted by Professor Imai is systematically biased in favor of finding spurious effects.

The following section presents the corrected estimates from Gerber and Green 2000. The new version of the New Haven dataset shows no signs of randomization failure and generates statistical results that are substantively indistinguishable from what was reported in Gerber and Green 2000. Face-to-face visits increased turnout substantially; direct mail, modestly. Brief paid phone calls were ineffective. The section concludes with a replication of the New Haven study involving more than a million subjects. The results again show phone calls to be ineffective. The 6.5–percentage point effect claimed by Professor Imai (2005) is rejected at a significance level of less than one in a trillion.

The subsequent section explains why Professor Imai's statistical approach exaggerates the effectiveness of phone calls. We describe the faulty assumptions underlying his model and the nonstandard way in which he implemented propensity score matching. Replication of Professor Imai's (2005) analysis reveals a series of errors that call into question the magnitude, significance, and robustness of his matching estimates. This section concludes by showing that the biases of Professor Imai's matching analysis are not specific to the New Haven study; they can be reproduced using other experimental data.

Professor Imai raises a number of interesting and important methodological topics, ranging from research design to data analysis. However, this essay alerts readers to several instances in which Professor Imai's methodological advice is ungrounded in statistical theory and deviates from standard practice. The randomization check that he proposes produces incorrect p values and does not properly detect problems in the random assignment of experimental groups. His assurance that propensity score matching reliably estimates treatment effects is based on a dated and selective reading of the relevant literature, and at no point does he establish, analytically or empirically, the superiority of matching over other estimation methods, either in general or in this application. We show also that Professor Imai's advice about how to compare the cost-effectiveness of alternative get-out-the-vote tactics rests on an algebra mistake; using the correct formula confirms our original conclusions. His criticism of factorial design is groundless, both empirically and theoretically. His essay repeatedly points out the importance of assorted interactions and nonlinearities yet neglects to subject any of them to significance tests; doing so reveals that they are all insignificant. In sum, not one of the key substantive or methodological claims of Professor Imai's essay survives scrutiny.

CORRECTION OF PUBLISHED RESULTS

Returning to the original data files, we reconstructed both the household-level and the individual-level New Haven study for purposes of preparing this correction. We corrected errors associated with (1) imperfect matches among names returned by the phone bank, names returned by canvassers, and names on the original master file, (2) a miscommunication between us and the phone bank about which treatment groups were to be assigned a get-out-the-vote appeal, and (3) data manipulation errors that caused some people in the control group to be incorrectly labeled as treatment subjects.[1] The reconstructed dataset

Alan S. Gerber is Professor, Institution for Social and Policy Studies, 77 Prospect Street, P.O. Box 208209, Yale University, New Haven, CT 06520-8209 (alan.gerber@yale.edu).

Donald P. Green is Professor, Institution for Social and Policy Studies, 77 Prospect Street, P.O. Box 208209, Yale University, New Haven, CT 06520-8209 (donald.green@yale.edu).

The authors are deeply indebted to Kevin Arceneaux, who helped with every facet of this project. They also wish to thank Jack Citrin, Mo Fiorina, John Geer, John Krasno, Jeff Lewis, Doug Rivers, Jas Sekhon, Jeff Smith, Petra Todd, and three anonymous reviewers who commented on earlier drafts, and Robert Gentleman and John Immel, for their technical help.

[1] The datasets that we prepared for Professor Imai in June 2002 addressed problems 2 and 3. The correction to problem 3 accounts

includes all of the observations in the original assignment, including those not found on the cross-off sheets supplied by the registrar of voters after the 1998 election. The corrected data file may be found at http://research.yale.edu/vote/NewHaven1998.html, and it allows the option of analyzing all 31,098 observations or the 29,811 observations with valid 1998 vote data. Since missing data are uncorrelated with the original random assignment, the decision of how to treat missing observations has trivial effects on the results, as demonstrated below. In this section, we analyze the corrected dataset; in the next section, when replicating Professor Imai's matching analyses, we use the earlier versions of the dataset that he analyzed.

Randomization Check

Before analyzing the data, we first demonstrate that nothing is amiss in the random assignment of subjects to treatment groups. Since the data were randomized at the level of the household, this randomization check must also be performed at the household level ($N = 23{,}450$). Performing the randomization check at the *individual* level, as Professor Imai does (his Table 6), provides grossly misleading results because individuals living in the same household share the same background characteristics, such as voting ward. Monte Carlo simulations reported by Arceneaux, Gerber, and Green (2004) demonstrate that even when the household-level assignments to mail, phone calls, and visits are generated at random, estimating the randomization check at the individual level incorrectly declares one or more of these randomizations flawed in more than 95% of the simulated experiments. Professor Imai cites McCullagh and Nelder (1989, 119) as an authority for his test, but they warn (118, assumption 1) that their result hinges on the assumption that the observations are independent. Due to this error, all of the *p*-values reported in his Table 6 are incorrect.

Using logistic regression, we regressed each experimental assignment—mail, phone, and face-to-face canvassing—on all of the background information available for each household: the ward of residence, the number of voters in the household, the age of each voter, a dummy variable for missing age information, dummies indicating whether each voter abstained in 1996, dummies indicating whether each voter voted in 1996 (the omitted category was not being on the voter rolls in 1996), and dummies indicating whether each voter was registered with one of the two major parties. This exercise generates chi-square statistics with 39 degrees of freedom. We do not follow Professor Imai's procedure of including more than 100 first-order interactions among these variables.[2] Monte Carlo simulation shows that even when applied to household-level data, his method of saturating the model with interactions leads to rejection of the null hypothesis of random assignment at more than twice the rate implied by the critical chi-square value used in his test (Arceneaux, Gerber, and Green 2004). This error further distorts the results presented in his Table 6.

Applying the proper randomization check to these data reveals that the *p*-values from these regressions are nonsignificant: .95 for mail, .78 for phone, and .39 for face-to-face canvassing. A superior way to assess randomization is to analyze assignment to all eight experimental permutations simultaneously. A multinomial logistic regression with 273 degrees of freedom produces a nonsignificant χ^2 of 261 ($p = .70$). (Employing Professor Imai's method of including all first-order interactions raises the number of degrees of freedom to 1,764 and decreases the *p*-value to .27.) After deleting households that did not have valid voting data in 1998, we obtain similar results ($N = 22{,}449$); the *p*-value for mail is .93; phone, .76; and face-to-face canvassing, .43. The multinomial regression produces a *p*-value of .60.[3] In sum, households were randomly assigned to experimental groups, and the randomization checks reveal no unexpected correlations between random assignment and voters' background characteristics.

Professor Imai (2005, 290) concedes that two-stage least-squares (2SLS) is the best estimator for gauging treatment effects when the treatments are assigned at random. This estimator is the multivariate generalization of the instrumental variables estimator proposed by Angrist, Imbens, and Rubin (1996). The vote is regressed on actual treatment using treatment assignment as an instrumental variable (see Statistical Implications of Failure to Treat, below). The next section applies this estimator to the corrected experimental data.

The Treatment Effects Are Substantively Identical to Those Reported in Gerber and Green 2000

Using the corrected individual-level data, Table 1 reports the 2SLS regression estimates. Looking first at the model without covariates, we see that the estimated effect of face-to-face canvassing changes from 8.7 (published) to 8.4 (corrected). Mail's effect changes from 0.58 (published) to 0.51 (corrected). The coefficient estimate for phone calls remains negative. The published value of −4.7 is corrected to be −1.9. A similar

for the change in the number of treatment group observations from 958 to 805 in his Tables 1 and 2 and explains the slight increase in the contact rate reported in his Table 4. Problem 1 did not become apparent until a year later, when, with the assistance of a database programmer, we reconstructed the dataset from the four original data files (the master file of treatment assignments, the names and contact information reported by the phone bank, vote histories, and results from the door-to-door canvass). This process also corrected errors in the party and age variables.

[2] In his footnote 18, Professor Imai (2003) advocates the use of even more interactions and higher-order terms than he was able to implement in this application. For an example of a properly conducted randomization check, see Ansolabehere and Iyengar 1995, 172.

[3] Nonsignificant *p*-values are also obtained for all the subsamples that Professor Imai analyzes: $p = .54$ when excluding households that received multiple treatments and $p = .66$ when further excluding households that received a placebo phone call asking for a blood donation.

TABLE 1. Published and Corrected Regression Results

	Without Covariates		With Covariates	
Independent Variable	Published Coefficient (SE)	Corrected Coefficient (Robust SE)	Published Coefficient (SE)	Corrected Coefficient (Robust SE)
Personal contact	8.7	8.4	9.8	9.0
	(2.6)	(2.6)	(2.2)	(2.2)
Direct mailings (0 to 3)	0.58	0.51	0.63	0.59
	(0.27)	(0.30)	(0.23)	(0.25)
Telephone contact	−4.7	−1.9	−3.5	−1.7
	(2.3)	(2.4)	(2.0)	(2.0)
N	29,380	29,435	29,380	29,435

Note: Dependent variable is voter turnout in 1998. Entries are 2SLS estimates, with standard errors in parentheses. Covariates include past voting in 1996, registration as a Democrat or Republican, age, age squared, number of registered voters living in the household, and dummy variables marking the ward of residence. Standard errors for the corrected estimates take into account the clustering of individuals within two-voter households. The first-stage equations include variables indicating intent to treat with personal contact, direct mail, and telephone. The first-stage equation also includes covariates, when these are used in the second-stage equation. The second-stage equation includes the number of mailings sent, a dummy variable indicating whether any member of the household was contacted face-to-face, and a dummy variable indicating whether any member of the household was contacted by phone. Both the published and the corrected regressions delete observations with missing data on voter turnout in 1998 as well as observations with missing data for age. Note that cases with missing data for age are included in Table A1. The version of the data analyzed by Imai (2005) generates estimates that are almost identical to the correct coefficients in this table; for phone calls, the estimated treatment effect is −2.0 without covariates and −2.2 with covariates.

pattern holds when we estimate the treatment effects controlling for voting history, age, party registration, and ward of residence. Again, the published estimates change from 9.8 to 9.0 (face-to-face canvassing), from 0.63 to 0.59 (mail), and from −3.5 to −1.7 (phone). The inclusion of covariates reduces the standard errors slightly but otherwise has little effect on the results.

Correcting the coding errors in the New Haven study changes none of the substantive conclusions that flow from it. Door-to-door canvassing's effects are large and statistically significant using a one-tailed test. Mail's effects are small and borderline significant using a one-tailed test. Phone's estimates are negative and, of course, not significantly greater than zero using a one-tailed test.

The results do not change when one discards the experiment's factorial design. Professor Imai sharply criticizes factorial designs, referring to them in the text and tables as "incorrect." (This usage has the unfortunate consequence of creating confusion between our data-processing errors and the supposed "error" of assigning observations to multiple treatments.) His approach is to throw out all observations that were assigned to more than one type of voter mobilization treatment. We have more to say about this practice below. For the moment, note that the estimated effects of the phone treatment do not change substantively when one restricts the sample to this subset of observations, although the standard errors grow considerably because the *N* in the phone treatment group (excluding those assigned to the placebo condition) drops from 6,562 to 815. The phone treatment effect becomes −2.0 (SE = 6.2); with covariates, the estimate is −5.0 (SE = 5.3). Professor Imai leaves the impression that the negative estimates produced by the New Haven study reflect randomization failures or the misguided use of a factorial design. It is now clear that *the negative estimates have nothing to do with randomization failure or use of multiple treatments*. Table A1 (Appendix) shows that, regardless of sample definition or model specification, *all* of the models produce negative estimates. Some estimates are more negative than the results reported in Gerber and Green 2000.

What substantive inference do we draw from these negative estimates? Notwithstanding the title Professor Imai selected for his essay, at no point did we assert that get-out-the-vote phone calls reduce turnout. What we said (Gerber and Green 2000, 660) was, "Given our initial expectation that telephoning increases turnout, we take this to mean that the null hypothesis of no effect cannot be rejected using a one-tailed test."[4] On the first page of his essay, Professor Imai (2005) reports our conclusion this way:

> Gerber and Green (2000, 660) describe the negative effect of get-out-the-vote calls as "one of the most surprising results to emerge from our experiment."

But the passage that he quotes actually reads, "One of the most surprising results to emerge from our experiment is the ineffectiveness of telephone appeals."

Replication Confirms Original Results

We also conducted a replication study of sufficient size to potentially refute our earlier conclusions. This study

[4] Professor Imai relegates a fragment of this quotation to a footnote and inserts into his text a quotation from a paper on a different experiment published in another journal (Gerber and Green 2001). He quotes from a footnote in the 2001 article, where, after finding insignificant negative coefficients, we offer a speculation about why the commercial phone bank "conceivably" could have had a negative effect. The text of Gerber and Green 2001 (cf. 76, 80, 81, 82) leaves no ambiguity about our interpretation. We observed that "phone canvassing did not seem to affect voter turnout in our experiment" and that "get-out-the-vote calls produced no statistically significant increase in voting rates."

was designed to be above complaints about factorial designs or the inclusion of multiple individuals from the same household. Two national phone banks read a script much like the ones used in the 1998 study.[5] Note that this was a nonpartisan script, as distinct from the partisan appeals that Professor Imai (2005, 283) alludes to when bolstering the intuition that phone calls must work or else parties would not waste money on them. As in 1998, this script was read during the weekend leading up to the federal midterm elections.

The populations used for this study were registered voters in Iowa and Michigan. These states were chosen because they maintain high-quality computerized voter files with the vote history of each registered voter. We use this information later, when we apply matching to these data. The congressional districts of each state were divided into "competitive" and "uncompetitive" strata. Within each stratum, households containing one or two registered voters were randomly assigned to treatment and control groups.[6] Only one type of treatment was used: a get-out-the-vote phone call. Just one representative from each household was assigned to treatment or control; the other voter was ignored for purposes of calling and statistical analysis. Because only one member of each household was treated, no complications arise due to correlation within households. A total of 60,000 individuals were assigned to be called; the corresponding control group contains 1,846,885 individuals. At the time of this writing, voter turnout results for the November 2002 elections remain unavailable in two small Michigan counties. Eliminating those observations reduces the treatment group to 59,972 and the control group to 1,845,348.

Given the enormous sample size, the experiment can easily detect the 6.5–percentage point effect that Professor Imai reports in his Table 9. The 2002 results, however, closely resemble our 1998 findings. The effect of the treatment-on-the-treated, controlling for the design strata (competitiveness levels within each state), is 0.4 percentage point.[7] Due to the massive N, the standard error of this estimate is just 0.5, which means that the 99% confidence region extends from -0.9 to 1.7. The 6.5–percentage point effect generated by Professor Imai lies more than 12 standard errors away from the apparent parameter value. If the true parameter value were 6.5 percentage points, the probability of obtaining this experimental finding is less than 10^{-20}.

The experiments just summarized do not mean that all forms of phone canvassing are doomed to fail. As we stated (Gerber and Green 2000, 661), these findings demonstrate only that brief, mechanically delivered reminders to vote conducted by commercial phone banks have negligible effects on voter turnout. In subsequent research, we studied alternative phoning strategies to see how phone banking might be conducted more effectively. In Green and Gerber 2001, we drew the distinction between the commercial phone banks studied in Gerber and Green 2000, 2001 and volunteer phone banks delivering chatty, informal scripts. Ignoring this distinction (which is spelled out explicitly in the second paragraph), Professor Imai (2005, 298) cites this report as evidence that his "correction" squares with other results showing that phone banks increase turnout significantly. The thesis of Gerber and Green 2000 is not that door-to-door canvassing invariably works while phone calls invariably fail; it is that the effectiveness of alternative forms of voter mobilization increases when messages are delivered in an authentic, personal manner.

Cost Accounting. Professor Imai's central substantive claim is that we have overestimated the cost-effectiveness of personal visits vis-à-vis phone calls and direct mail. The experimental studies cited above show that calls from commercial phone banks have negligible effects on voter turnout. If four votes are produced for each 1000 people who are contacted by a commercial phone bank and if one pays 50 cents per contacted person, each additional vote costs $125. Commercial phone banks delivering brief nonpartisan get-out-the-vote calls are clearly not a cost-effective way to increase voter turnout (see Cardy 2005 and McNulty 2005 for similar findings regarding partisan calls from commercial phone banks). Let us now turn to Professor Imai's further claims concerning direct mail.

Professor Imai (2005, 290) contends that we have "underestimated the effectiveness of sending postcards by incorrectly comparing the estimated ITT [intent-to-treat] effects for postcards with the estimated treatment effects for visits."[8] Our mistaken conclusion, he asserts (289), is based on our alleged "assumption that all voters who were sent postcards actually received and read them." Such an assumption is not warranted because many cards may not have reached a voter due to changes of address or may have been discarded as "junk mail." We now review our calculations to demonstrate that they were correct and that Professor Imai's claim is based on an accounting fallacy.

To calculate the cost per vote of a mailing regimen, we reasoned as follows (Gerber and Green 2000, 661): "At fifty cents per mailer, sending three mailings to each household (containing an average of 1.5 voters) nets roughly one additional voter per $40 spent." This figure was calculated based on our two-stage probit

[5] The script read: "Hello, may I speak with [name of person] please? Hi. This is [caller's name] calling from Vote 2002, a non-partisan effort working to encourage citizens to vote. We just wanted to remind you that elections are being held this Tuesday. The success of our democracy depends on whether we exercise our right to vote or not, so we hope you'll come out and vote this Tuesday. Can I count on you to vote next Tuesday?"

[6] As a randomization check, we used logistic regression to predict treatment based on vote in 2000, age, number of registered voters in a household, and state house district. As expected, the chi-squares for each stratum are nonsignificant: Iowa noncompetitive ($df = 24$, $p = .49$), Iowa competitive ($df = 63$, $p = .72$), Michigan noncompetitive ($df = 95$, $p = .60$), and Michigan competitive ($df = 31$, $p = .23$).

[7] The effects of phone banks are similar across states (Iowa, $b = 0.6$, $SE = 0.6$; Michigan, $b = 0.1$, $SE = 0.8$), competitiveness strata (uncompetitive districts, $b = 1.0$, $SE = 0.7$; competitive districts, $b = -0.3$, $SE = 0.7$), and commercial phone bank firms (first phone bank, $b = 0.6$, $SE = 0.9$; second phone bank, $b = 0.2$, $SE = 0.6$). The results remain unchanged when controls are introduced for past voting behavior and age ($b = 0.4$, $SE = 0.4$).

[8] See below for a discussion of ITT effects.

estimates using the following formula:

$$\frac{\$1.50}{1.5 \cdot [\Phi(-0.138 + 3 \cdot 0.0214) - \Phi(-0.138)]}$$
$$\approx \frac{\$1.50}{1.5 \cdot 0.025} = \$40.00.$$

Using propensity score matching, Professor Imai reports in his Table 9 that three mailings increase turnout by 1.5 percentage points (SE = 1.0). This statistically insignificant estimate implies a cost-per-vote estimate that is even higher than ours: \$1.50/$(1.5 \times 0.015) = \$66.67$.

Next, we take up the charge that our comparison of the *relative* effectiveness of mail and visits is miscalculated. Here is how we calculated the cost-effectiveness of visits (Gerber and Green 2000, 661): "Similar calculations, using \$1.50 as the cost per personal contact (10 contacts per hour at \$15 per hour), produce an estimate of approximately one more voter for each \$8 spent." Expressed algebraically, that calculation looks like this:

$$\frac{\$15/10}{1.5 \cdot [\Phi(-0.138 + 0.323) - \Phi(-0.138)]}$$
$$\approx \frac{\$1.50}{1.5 \cdot 0.128} = \$7.81.$$

Professor Imai's matching estimate of 9.2 percentage points (see his Table 9) generates a cost-per-vote estimate of \$10.87 for personal visits. Either way, personal visits are more cost-effective than direct mail.

These calculations produce the same answers if we use ITT estimates and rescale the calculation to eliminate contact rates. Let C_a be the cost per *attempted* treatment, and let T_a be the number of votes produced per *attempted* treatment. The cost per vote of this intervention may be calculated as C_a/T_a. If, instead, we define C_r to be the cost of *actually* contacting someone, C_r is equal to C_a (the cost of attempting to treat) divided by α (the proportion of people in the treatment group who are actually treated). The number of votes produced per actual treatment is T_r, which equals T_a/α, since the treatment-on-treated effect is the intent-to-treat effect divided by the contact rate. It follows, therefore, that cost per vote equals

$$\frac{\textit{cost per treatment}}{\textit{votes per treatment}} = \frac{C_r}{T_r} = \frac{C_a/\alpha}{T_a/\alpha} = \frac{C_a}{T_a}$$
$$= \frac{\textit{cost per attempted treatment}}{\textit{votes per attempted treatment}}.$$

This equation refutes Professor Imai's (2005, 290) assertion that it is "inconsistent" to calculate cost per vote using treatment effects for visits and ITT effects for mail. Suppose we take Professor Imai's advice to allow for the fact that some people do not read their mail. That means dividing both the numerator and the denominator by α, which leaves the estimated cost-efficiency unchanged.[9] In sum, Professor Imai's argument about cost-efficiency reflects confusion about how to manipulate the treatment and ITT effects. When this confusion is cleared up, the cost-efficiency of personal visits remains superior to direct mail, regardless of whether one uses his estimates or ours.

WHY DOES PROPENSITY SCORE MATCHING PRODUCE INACCURATE ESTIMATES?

The experimental evidence leaves little doubt about whether brief nonpartisan phone calls from telemarketing firms are effective. The remaining question is why Professor Imai's matching analysis produced inaccurate estimates. This section traces the bias in Professor Imai's estimator to two sources, the assumptions underlying the method and the problematic way in which it was implemented and reported. First, we model the selection problem that this matching procedure purports to solve. Second, we note several instances in which Professor Imai exaggerates the inadequacies of instrumental variables estimation as a method for analyzing experimental data. Third, we show that the matching estimator that he offers as an alternative relies on much stronger substantive assumptions. Fourth, a review of the matching literature reveals that the general properties Professor Imai ascribes to this method are not supported by the empirical track record of this technique. Fifth, we show that the evidence that Professor Imai marshals to demonstrate the significance and robustness of his estimates is marred by computational errors and discrepancies between the results of his computer programs and what he reports. When these defects are corrected, it turns out that his method fails both of the specification tests that he proposes as ways of detecting bias in propensity score matching, as well as two further specification tests that follow from the logic of his model. Finally, we demonstrate that the biases of Professor Imai's estimator are not specific to the New Haven study. Applying matching to the replication experiments conducted in 2002 shows that this estimator again severely overestimates the effects of brief nonpartisan phone calls.

Statistical Implications of Failure to Treat

When estimating experimental treatment effects, analysts must attend to complications that arise when some of the people who are assigned to the treatment are not actually treated. In the case of the New Haven experiment, for example, 32% of the 5,275 households assigned to receive get-out-the-vote phone calls in fact received them. The remaining 68% had unknown or nonworking telephone numbers, never answered the phone, or hung up on callers before they could read their scripts. Failure to treat a portion of the assigned

[9] For example, if one believes that only one-fourth of those who were sent mail read it, the treatment-on-treated effect for mail would be four times greater than the ITT effect. However, the cost per actual treatment would also be four times higher than the cost per attempted treatment.

treatment group creates a selection problem: if those reached by phone have a greater propensity to vote than those who were not, a comparison of the treated group (those actually reached by phone) and the control group will overstate the effects of phone calls. Even if the phone calls truly have no effect on turnout, they may appear to have large effects.

In Gerber and Green 2000, we presented a model of the selection process and used it to derive an estimator of the treatment effect that is consistent even in the presence of this selection problem. Suppose that for a given canvassing effort, the population can be divided into two groups, those who are reachable and those who are not. Let α be the proportion of the population that is reachable. Let p_{nr} be the probability that a nonreachable person votes. Let p_r be the probability that a reachable person votes in the absence of an experimental treatment, and let $p_r + t$ be the probability that a reachable person votes after exposure to the experimental treatment.[10]

In the control group, we do not observe who is reachable and who is not. What we observe is the voting rate for the group as a whole. The expected voting rate in the control group (P_C) is

$$P_C = \alpha p_r + (1 - \alpha)p_{nr}. \quad \textbf{(1)}$$

When treatment and control groups are formed randomly, both groups have the same expected proportions of reachable and nonreachable people. In this case, the expected voting rate in the treatment group is

$$P_T = \alpha(p_r + t) + (1 - \alpha)p_{nr}. \quad \textbf{(2)}$$

Manipulating these equations and solving for t suggests the instrumental variables (IV) estimator for the treatment effect:

$$\hat{t} = \frac{\hat{P}_T - \hat{P}_C}{\hat{\alpha}}. \quad \textbf{(3)}$$

The numerator of this equation is the estimated ITT effect: it is simply the difference between the voting rate in the assigned treatment group (the subjects the experimenter *intended* to treat) and the voting rate in the control group. The denominator contains the "contact rate": the fraction of people in the treatment group who were actually treated. Both quantities are easily calculated from experimental data. Because the contact rate is positive, ITT and t always share the same sign. We show below that Professor Imai's matching analysis does not satisfy this logical requirement. When analyzing the effects of phone calls, matching generates a negative estimate of the ITT but a positive estimate of the treatment effect.

Professor Imai concedes that when experimental groups are assigned randomly, IV is the appropriate estimator of the treatment-on-treated effect (often referred to as simply the "treatment effect"). He argues that when groups are not assigned randomly, the IV estimator generates biased results. This argument is moot with respect to the experimental evidence presented above; the updated version of the New Haven dataset, as well as the Iowa and Michigan replication studies, shows no signs of randomization problems. This is not the only concern that Professor Imai raises about IV, however. To justify matching, Professor Imai advances a number of additional criticisms of IV estimation. These criticisms rest on three flawed arguments.

i. The Small Sample Bias of the IV Estimator Is Irrelevant in This Application. Professor Imai (2005) writes that it is "important to note" the finite sample bias of IV estimation and claims that the small size of our treatment groups "suggests the importance of finite sample considerations" (his footnote 18). However, the "importance" of small sample bias in our particular case is merely an assertion. It is easy to assess the small sample bias, under various contact rates, using simulation methods. Doing so reveals Professor Imai's concerns to be groundless (for Monte Carlo results, see Gerber and Green 2004b).

ii. IV Regression Is Not "Inefficient." In his footnote 18, Professor Imai contends that IV is "inefficient" and cites Bound, Jaeger, and Baker (1995) for support. Bound et al. make no such claim. This appears to be one of several instances where the term of art "inefficient" is misused as a synonym for "large standard error." Professor Imai attempts to bolster this theoretical claim with evidence, repeatedly assailing the IV estimator on the grounds that its estimates are sometimes positive and sometimes negative when applied to the New Haven study. This line of argumentation conflates the properties of estimators (algorithms) with the behavior of estimates (statistics obtained from a particular sample). Professor Imai fails to explain that the apparent volatility of his IV estimates stems not from the properties of the IV estimator but rather from the simple fact that he restricts his attention to small subsamples of the New Haven data. To produce the astonishing −27 estimate that he twice mentions on the first page of his article, Professor Imai takes the subjects who were assigned to the phone treatment, discards those who were also assigned to receive mail or visits, and then discards those who reside in households with two registered voters. That leaves 7% of the original treatment group. It is hardly surprising that dramatic reductions in sample size increase the volatility of the IV estimates.

iii. Randomization Failure Does Not Necessarily Lead to Bias in IV Estimation. In the model presented above, randomization ensures that both the treatment and the control groups have the same expected proportions (α) of reachable and nonreachable people. In any given dataset, these proportions may differ due to sampling error. Professor Imai's discussion of randomization frequently confuses the properties of randomization as a procedure with the outcome that results from a given random assignment. Randomization

[10] The parameter t can vary across those contacted, in which case t is interpreted as the average treatment effect among those actually treated.

is an unbiased procedure even if in a particular experiment it produces a treatment group with too many young people or too many people in a certain section of town. To the extent that randomization generates *observable* differences between treatment and control groups, an appropriate corrective is to control for these differences. For this reason, Gerber and Green 2000 reports the experimental estimates both with and without controls for background characteristics. This point is glossed over in the "replication" tables presented in Imai 2005, which exclude covariates.

How is the IV estimator affected when randomization is compromised by clerical error? Suppose that certain subjects had a greater probability of being assigned to the treatment group. This failing is not a sufficient cause of bias. Bias results only when the experimental assignment is correlated with omitted causes of the vote. Thus, even if Professor Imai had obtained significant results from a properly performed randomization check, he would not have established that the systematic patterns in the experimental assignment produced biased estimates. Professor Imai attempts to fill this logical gap by demonstrating that the treatment and control groups have different propensities to vote. He reports that in *select subsamples*, there are small and marginally significant correlations with past voting (Imai 2005, 293). Professor Imai neglects to mention that none of the treatment groups (phone, mail, or visit) bears a significant relationship to past voting in the sample as a whole. Nor are the treatment assignments correlated with age, number of voters in the household, or party registration. The largest absolute correlation between the phone treatment and any of these variables is .01 (nonsignificant), and jointly they do not significantly predict any treatment assignment in a logistic regression. It is not surprising, therefore, that correcting the clerical errors in the original New Haven data had little effect on the results.

We now turn our attention to propensity score matching, which Professor Imai represents as a superior statistical method.

Matching Theory and Implementation

When Is The Matching Estimator Unbiased? IV compares the turnout rates of those *assigned* to the treatment group with the turnout rates of those assigned to the control group, because these groups have been formed in a way that ensures their comparability. The point of performing randomized experiments is to liberate analysts from making strong assumptions about the unknown values of p_r and p_{nr}.

Professor Imai's matching analysis, by contrast, compares those who actually received the treatment to the control group, after controlling for background characteristics. (This type of comparison has been shown to produce biased estimates in fields such as biostatistics; see Lee, Ellenberg, Hirtz, and Nelson 1991.) Suppose one were to restrict the sample to subjects who share exactly the same background characteristics, such as age, party, and voting history. For this subgroup, Imai's estimator may be expressed as

$$t^* = p_r + t - P_C = p_r + t - \alpha p_r - (1-\alpha)p_{nr}$$
$$= t + (1-\alpha)(p_r - p_{nr}). \quad \textbf{(4)}$$

The last term in this equation represents the bias in the matching estimator. When $p_r = p_{nr}$, this last term cancels, and the bias is zero. In other words, the matching approach will work if, after controlling for background characteristics, reachable and nonreachable people become interchangeable. This is a very strong (and demonstrably false) assumption. Bear in mind that in the version of the New Haven data that Professor Imai analyzed, the voting rate among those reached by phone was 64.8%, compared to 44.4% among those in the control group. By controlling for background characteristics, Professor Imai whittles this 20.4–percentage point gap down to 6.5 percentage points. The question is whether the background information that he used was sufficient to control for all of the unobserved factors that could account for higher voting rates among those who answered the phone. The experimental results as well as the specification tests presented below demonstrate that the answer is no.

Are Matching Estimators Accurate in General, and Should We Expect Matching to Work in This Application? Because Professor Imai's estimator compares those actually treated to a comparison group of untreated individuals, it must overcome a selection problem. Although he declares that "when treatment assignment is not completely random and important covariates are observed, propensity score matching is the best available statistical method" (295), his enthusiasm is not shared even by authors of the articles that he cites. Heckman et al. (1998, 1019), for example, caution:

> In general, matching is not guaranteed to reduce bias and may increase it. . . . Moreover, matching is open to many of the same criticisms that have been directed against traditional econometric estimators because the method relies on arbitrary assumptions. Even with the rich data at our disposal, the method of matching is not, in general, an effective evaluation method.

Among recent essays that describe the array of choices available to investigators trying to measure treatment effects, none give pride of place to cross-sectional propensity score matching, the estimator that Professor Imai endorses (e.g., Angrist and Krueger 1999, Heckman, LaLonde, and Smith 1999, and Smith and Todd 2004). Also pertinent are meta-analyses by Bloom et al. (2002) and Glazerman, Levy, and Myers (2003), which evaluate various observational estimation techniques, including matching. They report that matching methods often fail to produce accurate treatment effect estimates.

Propensity score matching's mixed record is not reflected in Professor Imai's essay. He does not cite a single work on propensity scores written by others in the past six years, with the lone exception of Dehija and

Wahba (1999). Professor Imai (2005) cites this article to bolster the claim that "matching is known to effectively reduce bias caused by nonrandom treatment" (295) but neglects to report a well-known critique of their work, which claims that Dehija and Wahba's results were an artifact of questionable sample restrictions and faulty analysis (Smith and Todd 2001).

Are there any special reasons why propensity score matching is particularly well suited to gauging the effects of voter mobilization campaigns in the New Haven study? Professor Imai (2005, 295) offers three reasons to believe so:

> First, since the large control group roughly represents the population, we know that potential compliers exist in this group. Second, as shown later, I find many exact and close matches, indicating that the bias due to inexact matches is minimal. Third, when the covariates measuring important characteristics of the subjects are available, propensity score matching is a powerful method for reducing bias.

The first argument presupposes that the background information that Professor Imai has at his disposal (age, ward of residence, party registration, past voting behavior, and number of voters in the household) can be used to identify people in the control group who have the same probability of voting as the people who received phone calls. What about people who have moved? What about people with unlisted phone numbers? What about people who are unwilling to talk with phone canvassers? It is doubtful that these background variables can be used to find appropriate matches for those who actually receive the treatment. Bear in mind that the pseudo-R^2 of the propensity score model predicting who is treated is just .05.

Professor Imai's second argument avoids the central issue, which is the possibility that individuals who are successfully treated are different from individuals in the control group who share the same background characteristics. This argument also turns out to be misleading since the matching estimates reported in his Table 9 for the phone treatment are not based on exact matching. When estimating the effects of phone calls, Professor Imai matches each person who received a call to five people in the control group with the nearest propensity scores. Only a small proportion of these matches is exact. As we show below, the problem of inexact matches contributes to the bias in Professor Imai's estimates.

Professor Imai's third argument repeats his unsupported assertion about the power of matching to reduce bias in general. The analytic and empirical literatures on the subject give no such assurances about the general properties of the method.

Problematic Implementation of Propensity Score Matching. Table 2 shows that Professor Imai also exaggerates the magnitude, significance, and robustness of his matching estimates. We summarize four errors in Professor Imai's text and tables; interested readers may replicate these results for themselves using the replication programs we posted.

Error 1: Using Bootstrapping to Provide Point Estimates, as Opposed to Sampling Variances. Bootstrapping is "a tool for developing inferential statistics (i.e., confidence intervals and bias estimators), not point estimators of parameters" (Mooney and Duval 1993, 60). The point estimates that Professor Imai reports in his Table 9, however, were obtained by drawing N observations with replacement from the actual sample and then averaging over 500 replications of this procedure. This bootstrapping procedure generates distorted results. Table 2 compares bootstrapped estimates with estimates from the actual sample. The estimated effects of phone calls fall to insignificance when based on the actual sample. His matching program generates an estimate of 5.1 (SE = 3.3) when performing one-to-five matching on the actual sample, and this number drops to 3.6 (SE = 4.5) when one-to-five matching is performed with replacement.

Error 2: The Reported Bootstrapped Estimates in Professor Imai's Table 9 Are Outliers. Table 9 of Imai 2005 exaggerates the significance of the phone effect by reporting a coefficient of 6.5 with a standard error of 3.2. These numbers imply a t-ratio of 2.03, which tells the reader at a glance that the effect is statistically significant using a two-sided test ($p < .05$). When these numbers are replicated using 500,000 bootstrap samples rather than 500, the correct estimate is 6.2, with a t-ratio of 1.87 (see Table 2). When we repeated Professor Imai's procedure of 500 bootstrap samples 1,000 times, only 2% of the time did we obtain a t-ratio as high as 2.03.

Error 3: Imai Misreports the Results of His Two "Sensitivity" Tests. Professor Imai describes two sensitivity tests for his implementation of one-to-five matching. The first test indicates that the results do not depend on whether one performs one-to-five matching or one-to-one matching. Acknowledging that one-to-many matching causes the match quality to deteriorate, Professor Imai (2005, 299) notes, "If the results based on one-to-many matching are significantly different from those of one-to-one matching . . . we may conclude that the former suffers from large bias." His Table 9 indicates that one-to-five matching produces an estimate of 6.5; his text reports that one-to-one matching generates an estimate of "six" (299). This is incorrect. The one-to-one matching estimate is in fact 4.3, with a standard error of 4.1. Moreover, as the first row in Table 2 indicates, there is a clear relationship between the number of matches and the size of the estimated treatment effect: the greater the number of matches, the larger the estimated effect. The second sensitivity test is found in Professor Imai's footnote 24: "I report the results for matching without replacement, but the sensitivity analysis using matching with replacement produced similar results." The exact numbers behind these "similar results" are not reported. We report them in Table 2. Using Professor Imai's program, we find that the bootstrap estimate for one-to-five matching with replacement is 4.0 (SE = 4.5).

TABLE 2. Replicating Professor Imai's Matching Results Assessing the Effects of Phone Calls

	Number of Matches for Each Treated Observation				
Effect	1–1	1–2	1–3[a]	1–4	1–5
Treatment					
Bootstrap sample[b]					
Without replacement	4.3[c]	4.8	5.2	5.8	6.2[d]
	(4.1)	(3.7)	(3.5)	(3.4)	(3.3)
With replacement	3.7[e]	4.1	3.8	3.8	4.0
	(4.8)	(4.6)	(4.4)	(4.5)	(4.5)
Actual sample[f]					
Without replacement	2.9	4.1	4.5	4.6	5.1
	(4.1)	(3.7)	(3.6)	(3.4)	(3.3)
With replacement	3.6	3.5	3.6	3.6	3.6
	(4.8)	(4.6)	(4.4)	(4.5)	(4.5)
Intent to treat					
Bootstrap sample[b]					
Without replacement	−2.7	−2.8	−2.7	−2.6	−2.4
	(2.4)	(2.2)	(2.1)	(2.0)	(1.9)
With replacement	−2.7	−2.6	−2.6	−2.7	−2.7
	(2.6)	(2.6)	(2.6)	(2.6)	(2.6)
Actual sample[f]					
Without replacement	−2.4	−2.8	−2.8	−2.1	−2.5
	(2.4)	(2.2)	(2.1)	(2.0)	(1.9)
With replacement	−2.1	−2.2	−2.2	−2.2	−2.2
	(2.6)	(2.6)	(2.6)	(2.6)	(2.6)

Note: The propensity scores used for matching were generated based on the specification that Professor Imai (2005) reports in his footnote 21. Professor Imai verified via e-mail correspondence (September 19, 2003) that this was the specification used to generate his 1–1 and 1–5 matching estimates.

[a] Imai 2002, (Table 7) reports 1–1 and 1–3 matching but not 1–5 matching. Imai 2005, (Table 9) introduces 1–5 matching and omits 1–1 and 1–3 matching.

[b] Estimates are averages of 25,000 bootstrapped samples. Note that using bootstrapping to obtain point estimates deviates from standard statistical practice but is used here to replicate Imai's results.

[c] Imai (2005, 299) reports this estimate as 6 (standard error not reported).

[d] Imai (2005, Table 9) reports this number as 6.5, with a standard error of 3.2. Our estimate is based on 500,000 bootstrap samples.

[e] Imai (2005, ftnt 24) describes this number as "similar" to 6 (standard error not reported).

[f] This analysis does not draw bootstrap samples from the actual data. The only random element in this analysis concerns the breaking of ties between equally good matches. These results are based on the average estimates over 100 replications. The standard errors are the bootstrapped standard errors reported in the rows above.

For one-to-one matching, the figure is 3.7 (SE = 4.8). These numbers convey a different impression of the treatment effect from the estimate of 6.5 (SE = 3.2) reported in his Table 9.

Error 4: A Computer Error Accounts for the Supposed Robustness of Matching Estimates. A recurrent theme of Professor Imai's paper is the superior performance of matching when applied to subsamples of the New Haven data, such as those living in households with either one or two registered voters. His argument is that a good estimator produces similar results across subgroups.[11] According to his Table 9, phone calls increased turnout by 6.9 percentage points in single-voter households and 6.1 percentage points in two-voter households. This pattern of similar estimates, however, occurs because Professor Imai's computer program indexes household size incorrectly. The correct estimates are 2.1 and 10.0.[12]

Two important errors of omission must also be mentioned. The first concerns the negative ITT effects of phone calls. Logically, the effect of assigning people to be called must be the same sign as the effect of actually calling them. Although Professor Imai reports the ITT estimates using regression and claims that ITT estimates are needed for cost-effectiveness comparisons, he reports the matching estimates of the ITT effect only for mail, not for phone calls. Instead, the reader is shown only the positive matching estimates for the phone treatment effects and is informed that the

[11] This hypothesis shifted over successive versions of Professor Imai's *APSR* article. Professor Imai (2002, 25) earlier predicted a larger effect for one-voter households. The matching estimates reported in that paper supported his prediction.

[12] When reporting the corrected results, we have followed Professor Imai's practice of dividing the sample according to whether the *treated* subjects live in one- or two-person households. Professor Imai's program matches these subjects to the nearest propensity scores in the control group, regardless of whether these control subjects reside in one- or two-person households. He does not alert the reader to this fact, which invalidates the comparison between matching and IV for one- and two-voter subgroups. The computer error does not affect his estimates for mail. Without explanation, he estimated these matching models separately for one- and two-person households.

negative estimates generated by OLS and 2SLS somehow reflect the deficiencies of these estimators when applied to flawed experimental data. Table 2 shows that the negative ITT estimates have nothing to do with the use of linear, parametric estimators. Using Professor Imai's matching program, we estimated the ITT effect of phone calls. Regardless of the number of matches and regardless of whether one uses the actual sample or bootstrap samples, propensity score matching always produces insignificant *negative* ITT estimates, ranging from −2.8 to −2.1, with standard errors ranging from 1.9 to 2.6. (Bowers and Hansen 2005, analyzing the same data, also obtain weakly negative ITT estimates using matching.) The implication is that the effects of phone calls were statistically indistinguishable from zero.

The second omission is the failure to check whether the nontreated vote at the same rate as the control group. The key identifying assumption in Professor Imai's estimator is that the people who are reached by phone (the "treated" group) are identical to the people in the control group who share background characteristics. This reasoning leads him to infer the causal effect of phone calls from the fact that these two groups vote at different rates. Looking only at Professor Imai's results, one cannot know whether the observed differences between the treated group and the matched control group reflect a true positive effect of the treatment or the upward bias of the matching procedure (see Eq. [4] above). But the same logic also implies that people who were assigned to the treatment group but *not* contacted will vote at the same rate as those in the control group with similar background characteristics (see Imai 2005, Figure 1). Since there can be no treatment effect without a treatment, the difference between the untreated and their matched counterparts can only be attributed to the bias of the matching estimator.

Using Professor Imai's matching program to compare the noncontacted members of the treatment group to the control group reveals that those who could not be reached by phone vote at substantially lower rates. The estimate produced by one-to-five propensity score matching is −5.6, with a standard error of 2.3.[13] The hypothesis that matching produces unbiased estimates is therefore easily rejected at the 5% level. Note that the negative matching estimate implies that placing phone numbers on a list and *not* calling them depresses turnout. The absurdity of this proposition makes it a powerful diagnostic tool. It demonstrates that unobserved causes of voter turnout are correlated with whether a subject is reachable by phone.

The remaining question is whether the pattern of biases detected in the New Haven data can be reproduced in other studies. The answer is yes. When we apply matching to the enormous Iowa and Michigan samples, using covariates similar to those available in the New Haven study, we obtain upwardly biased estimates of 4.2 (SE = 0.4) and 3.2 (SE = 0.5), respectively, for the effects of phone calls.[14] We also find strong evidence that those who were not reachable by phone were less likely to vote than the control group, a clear sign of bias. In sum, matching produces misleading estimates in this application.

COMMENTS ON INTERPRETATION AND RESEARCH DESIGN

Professor Imai contends that, when properly analyzed, the New Haven experiment shows that phone calls dramatically increase voter turnout. Due to the ready availability of experimental data, Professor Imai's conclusion can be evaluated with rare precision. The New Haven experiment and other large-scale field experiments demonstrate the ineffectiveness of brief calls from commercial phone banks. The probability that the treatment effect is as large as the 6.5 that Professor Imai presents is effectively zero.

Having addressed Professor Imai's main statistical claim, we conclude by briefly responding to three further criticisms that he levies against the analysis presented in Gerber and Green 2000. These criticisms allege that (1) the use of factorial design in the New Haven experiment was "incorrect" and "inefficient," (2) we failed to take notice of many important nonlinearities and interactions, and (3) we resist the use of advanced statistical methods.

Factorial Design

Professor Imai (287–288) sharply criticizes the use of factorial design and declares the use of multiple treatments "incorrect" and "inefficient." Asserting that "in principle, it is advisable to minimize the number of treatments in field experiments" (288), his correction to the New Haven study is to discard all observations assigned to multiple treatments (289). We would caution readers that Professor Imai's practice of discarding all observations assigned to multiple treatments is highly idiosyncratic. He cites no authorities, provides no statistical analysis, and ignores the large and trenchant literature that endorses factorial design, particularly in the early phases of a research program (see Mead 1988, 584–85).

Professor Imai's discussion of factorial design is predicated on a mistaken understanding of what it means to estimate the "marginal effect" of a treatment. This endeavor does not require one to compare a single

[13] The propensity score model used to generate this estimate uses all of the main effects and first-order interactions among the background variables. This model passes the balancing tests described by Professor Imai.

[14] The large sample sizes of these studies enable us to use exact matching, rather than propensity score matching (see Rosenbaum and Rubin 1985 on the superiority of exact matching). Because the treated and control observations share exactly the same values of the covariates, the two groups are automatically "balanced." In Iowa, 10,299 of the treated observations (75%) were matched to 132,577 observations in the control group. In Michigan, 9,565 of the treated (85%) were matched to 167,599 observations in the control group. To mimic the New Haven study, the control groups in Iowa and Michigan included randomly selected individuals without known telephone numbers.

treatment to a pure control group. On the contrary, a marginal effect refers to a change from one experimental condition to another, holding other factors constant. For example, to estimate the marginal effect of a visit, we might compare those who were assigned only visits to those who were assigned nothing. But we might also look at those who received visits and mail to those who received only mail. The latter simulates the effects of visits in a campaign environment in which voters are receiving mail. Neither can be deemed "correct" unless one stipulates that only certain kinds of campaign environments are worth studying. To characterize the design of the New Haven study as "inefficient" because it did not focus on just a single type of marginal effect fails to appreciate the aims of the experiment, which sought to explore treatment effects across a range of different conditions. And to denounce the use of factorial design because it requires the analyst to impose "assumptions" about the additivity of treatment effects fails to understand that these supposed "assumptions" are the questions that a factorial design addresses empirically.

In addition to these general misgivings, Professor Imai also marshals a specific substantive claim to justify discarding 89% of the observations that received the phone treatment. He argues that the effects of phone calls could be diminished among those receiving mail or personal visits (287). The New Haven data, however, provide no empirical support for this particular conjecture; the phone effects turn out to be slightly larger among those who received multiple treatments. In sum, Professor Imai's forceful rejection of the use of factorial design is groundless. For a further analysis of the statistical properties of Professor Imai's approach, see Gerber and Green 2004a.

Interaction Effects and Nonlinearities

The leitmotif of Imai 2005 is the importance of recognizing interactions and nonlinearities that might otherwise be ignored in a parametric statistical analysis. The reader is repeatedly alerted to nuances in the experimental data that were evidently ignored or concealed in Gerber and Green 2000. On page 290, Professor Imai asserts that turnout is a nonlinear function of the number of mailings (in particular, he contends that the effects of mail *increase* in the number of mailers sent, which contradicts the diminishing-returns rationale he advances when rejecting factorial design); on page 292, he argues that the effect of the phone treatment varies according to the message used; on page 298, he maintains that mail has more influence on two-person households than one-person households. Not one of these *ex post* hypotheses is subjected to a statistical test. Doing so reveals each of these interactions to be insignificant, even at a charitable .10 significance threshold.

The Importance of Statistical Methods

The broader thesis of Professor Imai's (2005) essay, reflected in its title, is that "statistical methods are essential to the analysis of field experiments" (299). We do not disagree with the bromide that statistical methods are essential. Indeed, Gerber and Green 2000 presents the first analysis of a voter turnout experiment to make explicit statistical allowance for failure to treat all members of the assigned treatment group. Prior experimental studies, the results of which Professor Imai cites to support his conclusions, produced biased statistical results because they ignored failures to treat or categorized untreated subjects as though they had been assigned to the control group.

Finding no resistance to statistical methods in Gerber and Green 2000, Professor Imai takes aim at a quotation from our *State of the Discipline* essay in which we point out that when data are generated using random assignment, "Rudimentary data analysis replaces scores of regressions, freeing the researcher from the scientific and moral hazards of data mining" (Green and Gerber 2002, 810—11). He dismisses our suggestion that analytic methods that follow directly from the experimental design might have important advantages.

The argument we advanced in our *State of the Discipline* essay stems not from an antiquarian taste for the simple but rather from a recognition that transparent statistical analysis limits the discretion that analysts have when extracting conclusions from data. The more opaque and complex the procedure, the more opportunities for the analyst's conscious or unconscious biases to intrude. Lest the reader think this merely a theoretical concern, consider the role of discretion in Professor Imai's matching analysis. Professor Imai faced a wide array of choices when devising his matching procedure: whether to match with or without replacement, how many matches to select for each person who received a treatment, and whether to report the results from the actual data or from samples simulated using bootstrapping procedures. As we have shown, each of these choices is consequential for the estimates he obtains.[15] Each choice had the effect of increasing the magnitude and significance of the apparent effect of phone calls.

By rejecting Professor Imai's methodological recommendations, we do not mean to signal a resistance to statistical innovation or advice. The design and analysis of the New Haven experiment left room for improvement. First, the fact that randomization occurred at the household level means that the standard errors associated with analysis at the individual level must be corrected for within-household correlation. The tables presented here make that correction, but our original paper did not. Second, the power of the study could

[15] This list could be expanded to include the specification of the propensity score model on which matching was based. The specifications Professor Imai employed were not obtained using a well-defined model selection algorithm, and multiple models satisfy the criteria he lists on page 296. For an example of a well-defined model selection algorithm that does not involve looking at the matching estimates, see Diamond and Sekhon 2005. Furthermore, the performance of matching in this application shows that the balancing statistics that he reports are not sufficient to ensure that matching produces accurate treatment effect estimates.

have been improved if we had randomized the phone treatment after identifying the population of households that had known telephone numbers. Instead, we randomly assigned the treatment group and obtained updated phone numbers for it. Some of these households were found to have no working number. Because these households could not be reached, the contact rate in the treatment group suffered, raising the standard errors of our estimates.

At a more basic level, our data contained clerical errors. Even though they did not affect our conclusions, we find these errors distressing and apologize for them. Professor Imai offers the consolation that "randomization of treatment assignment is not as easy to accomplish as one might expect" (293). Tempting as it may be for us to blame our mistakes on the inherent difficulty of randomizing treatment assignments, the fact is that randomization is a simple procedure. Maintaining the integrity of randomization after assignment is a matter of being alert to potential problems associated with data processing, such as merging administrative records from different sources.

APPENDIX

TABLE A1. Treatment Effects on Voter Turnout Under Varying Sample Definitions: New Haven Data (2SLS Estimates)

	b (Robust SE)			
	All Cases Included		Excluding Missing Vote Data[a]	
Independent Variable	Without Covariates	With Covariates	Without Covariates	With Covariates
Full sample				
Placebo cases included				
Personal contact	8.8	9.1	8.5	9.0
	(2.6)	(2.2)	(2.6)	(2.2)
Direct mailings (0 to 3)	0.51	0.58	0.55	0.62
	(0.30)	(0.25)	(0.30)	(0.25)
Telephone contact	−1.6	−1.1	−1.9	−1.5
	(2.4)	(2.0)	(2.4)	(2.0)
N	31,098	31,098	29,811	29,811
Placebo cases removed				
Personal contact	8.5	9.2	8.2	9.0
	(2.6)	(2.2)	(2.6)	(2.2)
Direct mailings (0 to 3)	0.62	0.68	0.67	0.72
	(0.30)	(0.25)	(0.31)	(0.26)
Telephone contact	−2.2	−1.5	−2.6	−2.0
	(2.4)	(2.0)	(2.4)	(2.0)
N	30,100	30,100	28,842	28,842
One-treatment sample				
Placebo cases included				
Personal contact	12.8	10.6	13.1	11.2
	(3.7)	(3.1)	(3.7)	(3.1)
Direct mailings (0 to 3)	0.72	0.61	0.80	0.69
	(0.37)	(0.31)	(0.38)	(0.31)
Telephone contact	−0.9	−3.2	−1.6	−4.7
	(6.0)	(5.1)	(6.2)	(5.3)
N	23,095	23,095	22,124	22,124
Placebo cases removed				
Personal contact	11.8	9.9	12.2	10.7
	(3.7)	(3.2)	(3.7)	(3.2)
Direct mailings (0 to 3)	0.77	0.65	0.87	0.75
	(0.38)	(0.32)	(0.39)	(0.32)
Telephone contact	−1.2	−3.5	−2.0	−5.0
	(6.1)	(5.1)	(6.2)	(5.3)
N	22,318	22,318	21,363	21,363

Note: Regressions based on "all cases" treat those with missing data for vote as having abstained. Regressions that include placebo cases treat those who received calls appealing for blood donations as part of the control group for the phone treatment. Regressions based on the "one treatment sample" exclude observations assigned to multiple treatments. For compactness, the table excludes coefficients for covariates (described in Table 1). Standard errors take into account the clustering of individuals within two-voter households.

[a] Missing data for vote arise from a combination of factors. Some of those on the registration list moved or reregistered. In three wards, a few pages were missing from the cross-off sheets provided by the registrar of voters. Missing data, however, are statistically independent of assignment to treatment and control groups.

REFERENCES

Angrist, Joshua D., Imbens, Guido W., and Donald B. Rubin. 1996. "Identification of Casual Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (June): 444–55.

Angrist, Joshua D., and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." In *Handbook of Labor Economics,* ed. Orley C. Ashenfelter and David Card. New York: Elsevier, 1277–1366.

Ansolabehere, Stephen, and Shanto Iyengar. 1995. *Going Negative: How Political Advertisements Shrink and Polarize the Electorate*. New York: Free Press.

Arceneaux, Kevin T., Alan S. Gerber, and Donald P. Green. 2004. "Monte Carlo Simulation of the Biases in Misspecified Randomization Checks." Manuscript. Institution for Social and Policy Studies, Yale University.

Bloom, Howard S., Charles Michalopoulos, Carolyn J. Hill, and Ying Lei. 2002. "Can Nonexperimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?" Working paper. Manpower Demonstration Research Corporation.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995." Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90: 443–50.

Bowers, Jake, and Ben Hansen. 2005. "Attributing Effects to a Get-Out-The-Vote Campaign Using Full Matching and Randomization Inference." Paper presented at the Annual Meeting of the Midwest Political Science Association, Chicago, Illinois, April 7–10.

Cardy, Emily Arthur. 2005. "An Experimental Field Study of the GOTV and Persuasion Effects of Partisan Direct Mail and Phone Calls." *Annals of American Academy of Political and Social Science.* Forthcoming.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94 (448): 1053–62.

Diamond, Alexis, and Jasjeet S. Sekhon. 2005. "Genetic Matching for Estimating Causal Effects: A New Method of Achieving Balance in Observational Studies." Paper presented at the Annual Meeting of the Midwest Political Science Association, Chicago, Illinois, April 7–10.

Gerber, Alan S., and Donald P. Green. 2000. "The Effects of Personal Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." *American Political Science Review* 94 (3): 653–64.

Gerber, Alan S., and Donald P. Green. 2001. "Do Phone Calls Increase Voter Turnout? A Field Experiment." *Public Opinion Quarterly* 65: 75–85.

Gerber, Alan S., and Donald P. Green. 2004a. "Note on the Conditions Under Which It Is Optimal to Discard Observations Assigned to Multiple Treatments in an Experiment Using a Factorial Design." Manuscript. Institution for Social and Policy Studies, Yale University.

Gerber, Alan S., and Donald P. Green. 2004b. "Monte Carlo Simulation of Small Sample Bias for Varying Contact Rates." Manuscript. Institution for Social and Policy Studies, Yale University.

Glazerman, Steven, Dan M. Levy, and David Myers. 2003. Nonexperimental Versus Experimental Estimates of Earnings Impacts. *The Annals of the American Academy of Political and Social Science*. 589 (1): 63–93.

Green, Donald P., and Alan S. Gerber. 2001. "Getting Out the Youth Vote: Results from Randomized Field Experiments." Unpublished report prepared for Pew Charitable Trusts, Yale University.

Green, Donald P., and Alan S. Gerber. 2002. "Reclaiming the Experimental Tradition in Political Science." In *Political Science: State of the Discipline*. Vol. 3, ed. I. Katznelson and H. V. Milner. New York: W. W. Norton.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 65 (2): 261–94.

Heckman, James J., Robert J. Lalonde, and Jeffrey A. Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs." In *Handbook of Labor Economics*, ed. Orley C. Ashenfelter and David Card. New York: Elsevier, 1865–2097.

Imai, Kosuke. 2002. "The Importance of Statistical Methodology for Analyzing Data from Field Experimentation: Evaluating Voter Mobilization Strategies." Manuscript.

Imai, Kosuke. 2003. "Essays on Political Methodology." Ph.D. dissertation, Department of Government, Harvard University.

Imai, Kosuke. 2005. "Do Get-Out-the-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments." *American Political Science Review,* 99 (2): 283–300.

Lee, Young Jack, Jonas H. Ellenberg, Deborah G. Hirtz, and Karin B. Nelson. "Analysis of Clinical Trials by Treatment Actually Received: Is It Really An Option?" 1991. *Statistics in Medicine.* 10(10): 1595–605.

Mead, R. 1988. *The Design of Experiments: Statistical Principles for Practical Application*. New York: Cambridge University Press.

McCullagh, Peter, and John A. Nelder. 1989. *Generalized Linear Models.* 2nd Ed. London: Chapman and Hall.

McNulty, John E. 2005. "Phone-Based GOTV: What's on the Line? Field Experiments with Varied Partisan Components, 2002–2003." *Annals of American Academy of Political and Social Science.* Forthcoming.

Mooney, Christopher Z., and Robert D. Duval. 1993. *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Newbury Park, CA: Sage.

Rosenbaum, Paul R., and Donald B. Rubin. 1985. The Bias Due to Incomplete Matching. *Biometrics*, 41 (1): 103–16.

Smith, Jeffrey, and Petra Todd. 2001. "Reconciling Conflicting Evidence on the Performance of Matching Estimators." *American Economic Review, Papers and Proceedings* 91 (2): 112–18.

Smith, Jeffrey, and Petra Todd. 2004. "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*. 125 (1–2): 305–53.